A Real-Time Online System for Analyzing Restricted Data fromthe U.S. National Center for Health Statistics' National Health Interview Survey

Gentleman, Jane F.
National Center for Health Statistics, Division of Health Interview Statistics
3311 Toledo Road
Hyattsville, Maryland 20782, U.S.A.
Email: JGentleman@cdc.gov

Abstract

The National Health Interview Survey (NHIS) has monitored the health of the United States noninstitutionalized civilian population since 1957. The survey is conducted by the National Center for Health Statistics (NCHS), which is the United States' official health statistics agency. NHIS data on a broad range of health topics are collected through personal household interviews. Public use microdata files are released annually. Data users can currently perform their own analyses of NHIS data by (1) downloading public use microdata from the NCHS Website; (2) accessing restricted microdata via the NCHS Research Data Center (RDC), under controlled conditions to protect data confidentiality; and (3) accessing preliminary quarterly microdata files available in the NCHS RDC during the nine months before release of the final microdata files. However, users of these options must program their own analyses. The University of Minnesota's Integrated Health Interview Series is another source of selected NHIS microdata, harmonized to facilitate time trend analysis. To provide additional services, NCHS is developing two real-time online analytic systems that will perform user-directed analyses of NHIS data. "System P" will analyze only public use microdata. "System R" will also analyze selected restricted NHIS variables. NCHS is currently working with a contractor to develop rigorous confidentiality screening methods for System R that will meet NCHS' high standards for disclosure avoidance. System R will utilize a cocktail of disclosure avoidance techniques and will focus on helping meet demands for state-level NHIS estimates. Currently, analysts must use the RDC to access state identifiers, which requires submitting a research proposal and paying a fee. System R will provide rapid, convenient access to state-level estimates and reduce and/or delay the need to use the RDC. This report will describe the benefits and challenges of developing System R and providing it to NHIS users.

The National Health Interview Survey

The National Health Interview Survey (NHIS) went into the field for the first time in July 1957. From the beginning, the survey was designed to represent the U.S. civilian noninstitutionalized population and serve a diverse community rather than focusing solely on selected policy or program needs. Topics presently covered by the relatively stable core of the survey include health status, utilization of health care services, health insurance coverage, health-related behaviors (such as use of tobacco and alcohol), risk factors, and demographic and socio-economic information. In addition, supplemental questions on special topics are added to the NHIS questionnaire each year, co-sponsored by government agencies other than the National Center for Health Statistics (NCHS).

The most recent extensive revision of the NHIS questionnaire was in 1997. Since then, the NHIS has collected data about all family members in the Family Section of the NHIS core, from one randomly selected adult (the "sample adult") in the Sample Adult Section, and about one randomly selected child (the "sample child") in the Sample Child Section. To improve precision of estimates for certain minority subpopulations, the NHIS has been oversampling black persons since 1985 and Hispanic persons since 1995. Also, since the NHIS sample was last redesigned in 2006, Asian persons have been oversampled, and the probability of selection as the sample adult has been increased for persons aged ≥65 who are Hispanic, black, or Asian.

The NHIS is in the field collecting data in face-to-face interviews virtually continuously throughout the year. Telephone follow-up is sometimes done to finish parts of the interview. The questionnaire is administered in either English or Spanish. The Census Bureau has been NCHS' contractor for fielding the NHIS since the inception of the survey. Each year, NCHS releases one year of NHIS microdata online in public use files that have been suitably scrutinized and processed to protect confidentiality. For a number of years, this data release has occurred less than six months after the end of the data collection year, an impressive accomplishment for a large, in-person survey. Paradata describing the NHIS interview process are released along with the annual public use files, and multiply-imputed income and earnings data are released about two months after the annual microdata release.

NHIS staff members analyze NHIS data and produce a variety of publications and presentations. In particular, the NHIS Early Release Program produces two quarterly reports (on 15 key indicators and on health insurance coverage), and one bi-annual report (on cell phone usage). To provide early access to microdata by outside analysts, the Early Release Program also produces periodic preliminary NHIS microdata files for use in the NCHS Research Data Center. The Early Release Program is so-named because its products are made available before the annual public use microdata files are released.

For more information about the NHIS, see:
http://www.cdc.gov/nchs/nhis.htm.
For a list of NHIS supplements and their co-sponsors, see:
http://www.cdc.gov/nchs/nhis/supplements_cosponsors.htm
For more information about the NHIS Early Release Program, see:
http://www.cdc.gov/nchs/nhis/releases.htm

System P and System R: Two online real-time systems for analysis of NHIS microdata

Users of the National Health Interview Survey have several resources for accessing the survey's microdata and analytic products. Users interested in a particular health subject or in health survey methods can consult appropriate reports and papers produced by NCHS analysts and by many others outside NCHS. Those who wish to conduct their own NHIS analyses can use the online NHIS microdata files, which are publicly released once per year, along with thorough documentation, and are available free of charge. They can also analyze the preliminary NHIS microdata files (described above) that are available in the NCHS RDC before the annual public release. Also, they can use microdata from the University of Minnesota's Integrated

Health Interview Series (IHIS), which is a collection of selected NHIS variables that have been "harmonized" to facilitate time trend analysis from the 1960s to the present. IHIS microdata are based on NHIS public use files and are available free of charge from the IHIS website at http://www.ihis.us/ihis/. The IHIS project provides extensive documentation and now has a tabulation capability. NHIS users whose analyses require access to restricted NHIS (or other NCHS) variables that are not released publically can use the NCHS Research Data Center (RDC), which provides on-site access facilities in Hyattsville, Maryland, in Atlanta, Georgia, and at Census Bureau RDCs across the country. The NCHS RDC also provides a capability for remote access to restricted variables. Analyses conducted via these RDCs are limited and are carefully supervised to protect confidentiality. RDC users must submit and have approved a project proposal, and fees are charged. See http://www.cdc.gov/rdc/ for more details about the NCHS RDC.

To provide additional mechanisms for analyzing NHIS data, NCHS is developing two online real-time analytic systems that will be publicly available without submission of project proposals and without charge:

- *System P* will provide analyses of the same public-use NHIS microdata files that are released online each year. ("P" stands for public use.)
- *System R* will provide analyses of public use NHIS microdata plus selected restricted variables, with a focus on state-specific analyses. (The NHIS does not currently release state identifiers on its public use files.) Analyses will not be "canned," but will be performed in real time, "on demand" from the user, but with strict screening for disclosure avoidance before provision of the analyses. ("R" stands for restricted variables.)

NCHS staff members are now working closely with an outside contractor on the following phases of the project to develop System R and System P:

1. We will develop, describe, and demonstrate the effectiveness of methods for screening analyses produced by System R that are sufficiently rigorous to meet NCHS' high standards for disclosure avoidance and confidentiality.
2. We will develop and demonstrate the usability of System R. We will produce appropriate documentation for use by those maintaining the system and those using the system. We will make plans for adding new capabilities and data to the system in the future.

   We will develop and demonstrate the usability of System P. We will produce appropriate documentation for use by those maintaining the system and those using the system. We will make plans for adding new capabilities and/or data to the system in the future.
3. The new systems will be installed and implemented at NCHS.

After implementation of System R and System P, the two systems will require ongoing maintenance. Additional analytic capabilities will be added to each system, as needed and as possible. Each new year of NHIS data will be added to both systems. New microdata to be added to System R will have to be processed to protect confidentiality. Disclosure avoidance techniques in System R will be monitored and adjusted as necessary.

These two new analytic systems will complement the mechanisms and opportunities already available for access to NHIS data. Many surveys already offer analytic capabilities like System P. Capabilities like System R are, however, very rare because of the complexity of real-time confidentiality screening for analyses, which is an ongoing area of research. The remainder of this paper will describe the expected benefits of System R and the challenges of planning, developing, implementing, and maintaining it.

Benefits of System R

The development of System R is motivated by the great and increasing demand in the United States for state-level analysis. Because health care is largely administered at the state level in the United States, state-level analysis is needed for research and for development and evaluation of health policies. In particular, monitoring the effects of the new U.S. Patient Protection and Affordable Care Act will require state-level analyses. To the extent permitted by the system's real-time confidentiality screening tests, users of System R will be able to obtain state-level estimates. If state-level estimates based on one year of NHIS data for a particular state do not meet confidentiality requirements, the system will attempt to provide two-year or three-year estimates.

Users of System R will be able to obtain their requested analyses virtually "on demand" (subject to passing the real-time confidentiality screening tests), without having to submit a research proposal and wait for it to be reviewed and approved, and without having to pay any user fees.

System R will provide assorted analyses (fewer than for System P because of disclosure avoidance requirements). The analytic "wish list" includes descriptive statistics; outlier identification; multi-way cross tabs; significance tests, standard errors, confidence intervals; assorted graphs (scatter plots, bar graphs, histograms, box plots, time trend graphs, U.S. maps displaying estimates by state, graphs able to show results for very large sample sizes effectively, etc.); crude rates; directly-standardized rates (using selected standard populations or user-provided standard populations); regressions; etc. These analytic capabilities will be implemented in stages.

The fact that analyses will be performed in real time (and not be "canned") will provide much more flexibility in what analyses System R can provide; the developers and maintainers of the system could not possibly anticipate and satisfy all possible user requests for analyses if analytic results had to be prepared ahead of time. Admittedly, however, providing only canned analyses would do away with the need for real-time confidentiality screening, thus simplifying (but certainly not eliminating) disclosure avoidance efforts.

System R will use proper methods for variance estimation that account for the complex sample design of the NHIS. Analyses of NHIS microdata dating back to 1997 (when major changes to the NHIS questionnaire were made) will be available to users of System R, and newly-available NHIS microdata will be added to the system each year.

Challenges of planning, developing, implementing, and maintaining System R

Disclosure avoidance

Proper confidentiality screening is the most critical and the most challenging requirement for System R. Analyses must be screened before the results are shown to the user. The user will not be permitted to see analytic results that do not pass the screening tests. Screening methods must be rigorous and sophisticated to meet the strict confidentiality mandates for NCHS data. A cocktail of disclosure avoidance methods custom-developed for System R will be used, some involving perturbation of the underlying NHIS microdata files, some involving real-time alteration of analytic results, and some resulting in refusal to provide the requested analysis. Microdata perturbation and analysis alteration methods currently being considered include subsetting (performing an analysis using a subset of the requested data), swapping (exchanging one person's data for another's), top-coding, combining variable categories, limiting cell sizes in cross-tabs, limiting the number of variables crossed in cross-tabs, rounding, providing no unweighted sample sizes, completely removing some variables from the underlying microdata files, etc. Limitations may have to be imposed on provision of national estimates to prevent ill-intentioned users from comparing System R analyses with analyses of public use files in order to learn more about System R's perturbation and alternation methods.

Special, complex methods will be used by System R to prevent "differencing attacks," in which ill-intentioned users try to gain information about inappropriately small groups of individuals by subtracting frequencies in one set of cross tabulations from frequencies in another.

Also to deter ill-intentioned users from attempting to "crack" the methods used for disclosure avoidance, a complex system to track the history of requested analyses will be implemented so that an analysis of a given "universe" utilizes the same perturbed set of data if the request is ever repeated.

Users will be informed that results from System R are not the same as what would be obtained from analyzing in-house NHIS data, but users will not be given details of perturbation and alternation techniques or their parameters.

Before going online, System R will have to be approved by the NCHS Disclosure Review Board.

System security

Once ready for public use, System R will have to undergo a series of approvals to ensure that system security is protected. This will take some time, as multiple levels of government have their own rigorous requirements for system security, and the developers of System R will have to demonstrate that System R meets those requirements.

Leaving the public use files and their timeliness unchanged

Confidentiality screening methods for System R must take into account the separate public availability of microdata from NHIS public use microdata files and linked files. NCHS has imposed non-trivial restrictions on System R that will require it to use confidentiality screening methods that will not alter or impose any concomitant new restrictions on the contents of public use NHIS microdata files or reduce the excellent timeliness with which those public use files are released.

Adequacy of state sample sizes

The usefulness of System R will be related to state-specific NHIS sample sizes. The NHIS was designed primarily to provide national and regional estimates. But the NHIS has PSUs in all states and the District of Columbia. State-level estimates using NHIS data are representative of their respective states, but because of sample size limitations, and depending on the specific estimates, some estimates may not have adequate precision to be useful. For example, NCHS is able to publish annual state-level estimates of health insurance coverage for the 20 largest of the 50 U.S. states. By combining data from multiple NHIS years, more state-level estimates with adequate precision can often be produced. In its next redesign, and sooner if possible, NCHS is hoping and planning to increase state sample sizes states to add precision to estimates and to add PSUs to some states to add breadth of coverage. The 2011 NHIS has an increased sample size within all states but the 18 largest states. In its next redesign (in 2014 or 2015), and sooner if possible, NCHS plans to add an address-based telephone component to the NHIS sample to further increase state-specific sample sizes.

For smaller states whose sample sizes do not, for confidentiality reasons, permit System R to provide estimates based on one year of data, it will be necessary for System R developers to pre-determine which specific non-overlapping two-year or three-year groupings will be made available for which state. Otherwise, a user could deduce results based on one year of data that were intended to be suppressed. For example, a user could perform a cross-tab using data for year t and year t+1 combined, then perform the same cross-tab using data for year t+1 and t+2 combined, after which the user could obtain the same cross-tab for data from year t+1 by calculating the differences between frequencies in the first two cross tabs.

Availability and quality of requested analyses

Tradeoffs are inevitable between disclosure avoidance and the availability and quality of analyses provided by System R. Users will become frustrated if they often request analyses that are then denied for reasons of confidentiality. Overly-perturbed microdata will yield inconsistent and/or unacceptably inaccurate analytic results. The developers of System R are currently perusing results of simulations in order to decide on the nature, variety, and extent of perturbations and alterations that will be imposed on the underlying microdata and analyses. Analyses of perturbed data and altered analyses are being compared to analyses of unperturbed data and unaltered analyses to help determine which cocktail of perturbation and alteration techniques will provided the needed balance between disclosure avoidance and analytic usefulness.

Retaining one-year age categories

Another non-trivial restriction that NCHS has imposed will require System R to permit age to be analyzed in one-year categories, and not allow forced collapsing of age into broader age groups. This is proving to be extremely challenging to System R developers, since age, in combination with other variables, is a good identifier of individuals. But NCHS recognizes that different age groups are needed to analyze different aspects of health, so analysts need to be able to form their own appropriate age groups. For example, the Affordable Care Act allows adult children aged 19-25 to be covered by their parents' health insurance policies, but age group 19-25 is a non-standard age group, and an analysis of persons in that age group could not be properly conducted if analyses were limited to using standard 5-year age groups.

Balancing the budget

U.S. Federal budgets are generally allocated on a year-to-year basis, and optimal use of the NHIS budget requires careful planning and flexibility. The first priority for use of the NHIS budget is to produce and disseminate high-quality, timely microdata files. After that, enhancing analytic opportunities for NHIS data users, such as by providing and maintaining online analytic systems, continues to be a very important goal for NCHS.

Future dependence on contractors and contractors' software

System R will use some software packages that are proprietary, such as the software that will produce variance estimates and the software that will perform data analyses. While that software is widely available commercially, its use in System R will reduce NCHS' freedom, once the system is implemented, to switch contractors or to choose to maintain the system itself.

Each year, after the annual release of NHIS public use microdata files, a new year of data will be added to System R. That will require proper pre-processing of the new year of data. The new data will need to undergo perturbation and testing for disclosure avoidance. If, after System R is implemented, NCHS decides to maintain System R itself, without assistance from a private contractor, NCHS will have to have experts on staff who are capable of performing the required perturbation and testing.

Concluding remarks

The National Center for Health Statistics is the nation's official health statistics agency, and the NCHS' National Health Interview Survey is the primary source of information on the health of the civilian, noninstitutionalized population of the United States. NCHS strives to meet the needs of its data users and thus to promote public health through providing valuable data and data products. NCHS is committed to confronting and meeting the challenges of developing and providing System R in order to reap the many benefits that such an analytic system will offer.